ECONSPEAK: A Journal of Advances in Management IT & Social Sciences,
Vol. 9, Issue 8, August 2019 Impact Factor 6.596 ISSN: (2231-4571)
www.skirec.org Email Id: skirec.org@gmail.com

BIG DATA ANALYTICS CHALLENGES AND SOLUTIONS BY HADOOP AND R DATA LANGUAGE

KUMAR GOURAV[#1], Dr. AMANPREET KAUR[#2]

[#1]Research Scholar, UIC, [# 2]Assistant Professor, UIC, Chandigarh University

Gharuan, Mohali, Chandigarh (India)

Abstract - Big Data has gained much momentum day by day in the digital and computing world. Big Data is an approach to accumulate the data having datasets that are large in size and that are analyzed very quickly. Data may be varied from structured to semi-structured that results inefficiently by using traditional methods used for the management of data. There are various sources and systems available to handle this data and for this various tools like HADOOP and R Data language are available. The aim of this study is to highlight some of the challenges and their solutions to overcome them by these techniques that are used for the implementation of Big Data Analytics.

HADOOP is a popular open source tool that implements the actions of managing and allocating the Big Data. To analyze Big Data is a challenging task because it contains large volume of data that should be scalable. To overcome these challenges Hadoop Ecosystem and Map Reduce techniques are used. The data when converted into knowledge, vision and understanding is known as data analytics that is the essential factor of statistics. To handle the data effectively and proper analysis of Big Data R data language is used. It is also an open source data analysis and programming language. The environment of R includes the set of tools that are ready to use and make it easy to manipulate the data, perform calculations and generate charts and graphs.

Keywords: Big Data, Big Data Analytics, HADOOP, R Data language

I INTRODUCTION

Big data is the data that is huge in size and also requires new technology to manage that data. Conventional technologies are not able to handle these massive amount of datasets for extracting useful information. Data is the dynamic source that is founded in different forms. The data that uses the predictive analysis, user behavior analytics and also extraction of data from the particular size of data set is known

as Big Data. Data sets increases frequently while gathering information-sensing Internet of Things (IoT) devices like wireless network sensors, software logs, identifying radio frequency readers, cameras microphones and mobile devices. There is rapid growth in the use of mobile phones and due to this usage huge amount of data is generated every second that is unable to be managed by using the conventional technologies. Now these days' big data is used in almost every field all over the world. Most commonly used areas are Social Networking Sites, Search Engines, Healthcare, Information Technologies and Stock Exchange. Due to its reliability it improves businesses and also creates long-term opportunities in each and every field.

Characteristics of Big Data

The following are the Characteristics of Big Data:

- **Volume**: The mass amount of data created within an organization. It is generated at uncertain rate ranging from petabytes to zeta bytes.
- **Variety**: The data that is created by individuals or by machines are in many forms, types and from various sources. Here the complexity of the data is also defined. The varied data may be structured, semi-structured and unstructured.
- **Velocity**: The speed of the data in which it is produced, processed and analyzed.
- **Veracity**: It Consist of the data that is uncertain and untrusted and generated by some hindered sources.
- **Visualization:** The important part is to represent this large data. Charts and graphs, complex spreadsheets and formulas are used to display the data.
- **Validity:** Accuracy and Correctness of data can be used.

Big Data Analytics

Peter Sondergaard said that "In 21st Century Information is the Oil and Analytics is the Combustion Engine." Big Data Analytics is the measure of transforming, observing and understanding the data with the objective of analyzing meaningful information, finding the proper result and taking the right decisions. Data and its analytics is the major part of the digital revolution. It is the combination of data mining and decision making. This technique is used in all sectors for increasing productivity and revenue in less and effective cost. The data is archived and produced by the individuals can be utilized in proper way only if it is analyzed properly. By the way of explanation, without using the proper analytics on the data, the data is just a resource but not an appropriate resource. This technique

checks this big amount of data to discover the hidden patterns, correlations and other insights. It also helps to analyze the data and get solutions. The information that is stored within the data identifies the data that is useful for decision making.

After the collection of data the process of analyzing of data is started. For analyzing of data various types of analytics are used for various types of data. The following are the different types of analytics that are performed on data.

- **Descriptive Analytics:** This technique is used to convert the big data to small bytes. It sorts the application and then generates various metrics that includes the monitoring of data in various process or multiple process. E-mails are used for monitoring the results. It is the most commonly used technique and preferred by number of organizations.
- **Predictive Analytics:** It is the technique that deals with the statistical modeling and also estimate the future possibilities. There are various factors used for establishing new statistical methods that deals with the big data like traditional statistical methods through which a small sample is taken into consideration and then result is correlated for the significance of a particular relationship. Computational efficiency method is used in which small samples that do not scale up to data and the distinctive feature that inherent the big data heterogeneity, correlations between data and noise accumulation
- **Prescriptive Analytics:** This technique is also known as a suggestion tool and used to resolve the cause-effect relationship between analytic results and optimization policies. It's result based on the feedback provided by predictive analytic models.
- **Diagnostic Analytics:** This technique is used to study the earlier knowledge and provides the reason how, what and why happened regarding the data. It is also used to discover the hidden patterns that helps to analyze the various elements that effects either directly or indirectly. This technique is frequently used in social media.

II LITERATURE REVIEW

According to Villars. R. L., 2011 Big Data is the information examination strategy that empowers the ongoing advances in advances that supports capturing of high-speed information, analysis of data and storage of data. Information sources reach out past the conventional corporate database to mobile phones data, e-mails and sensor-created data where the information is not limited to

organized database records but also to the unstructured information having no standard arranging.

Russom. P., 2011 clarified that big data investigates huge volumes of information that find the different actualities. Consequently, big data analytics is the progressed scientific system that is connected on huge informational indexes and uncovers and use business change.

Kubick. W.R, 2012 explained that big data has been applied to the datasets that develops so huge that they become cumbersome to work with utilizing conventional database the board frameworks. They are informational indexes whose size is past the capacity of generally utilized programming instruments and capacity frameworks to catch, store, oversee, just as procedure the information inside a middle of the road slipped by time.

According to Sutherland and Shan, 2014 big data is particularly depends on volume, speed and variety.

Adams. M.N., 2010 clarified the data analytics is the way toward applying set of calculations to examine sets of information and concentrate helpful and obscure the various patterns, connections and data.

As indicated by Song. Z. furthermore, Kusiak. A., 2009 big data analytics is utilized to extricate beforehand obscure, valuable and patterns of data from huge datasets just as to recognize connections among different put away factors. Henceforth, it significantly affected innovations and research then it became an advantage for the decision makers to learn from past information.

Bains. J.K., 2016 explained that information gathered put away and broke down in monstrous measures of information which is alluded as big data and analysis of that information is known as big data analytics.

As per Nada Elgendy, Ahmed Elragal, 2014 big data size range from couple of dozen terabyte to numerous petabytes in a solitary data sets. There are numerous challenges like to catch information, to analyze data, to store data, to share data and to visualize the data. There is a requirement for apply progressed expository methods on big data and this is the place where big data analytics makes the difference. The examination procedure is utilized to get past obscure, helpful shrouded designs, to remove valuable obscure connections.

P Harshawardhan, et. al, 2014 stated that Hadoop has dispersed handling power at a surprisingly minimal effort, making it a compelling supplement to a conventional endeavor information framework.

K. Shvachko, et. al, 2010 additionally featured that how big data is sorted out in present day disseminated file system explicitly the Hadoop Distributed File System (HDFS).

J. Dean and S. Ghemawat, 2008 explained that that how big data is prepared utilizing the MapReduce computational system.

J. Stanton, 2013 also highlighted that R data language is extremely helpful for analysis of big data as most of the big data is the collection of data in the form of text. It is efficient and effective language for data science. There are two modules of R that makes R to be an efficient data language. This combination is useful to use many methods of data science.

III TOOLS USED FOR BIG DATA ANALYTICS

There is rapid increase in the research and development in big data technologies. It includes various tools like computational frameworks, deployment strategies, algorithms, analytic platforms and additional services that makes a powerful big data ecosystem. The techniques that are used for analyzing big data have various approaches that are optimized for data-specific and domain-specific datasets. There are number of open source big data analytics tools are available. Here are some important big data tools.

HADOOP

It is an open-source software that is developed by using Java language and it can work with a block of data sets. Distributed storage processing is used for the big data sets and also the Map reduce programming model is used that is made up of clusters of computers. The Hadoop distributed file system (HDFS) contains the storage part in the core of apache Hadoop. It distributes the data into large blocks and then bundles the data so that it can process to large volume of data with different structures.

HADOOP Architecture

Hadoop architecture that provides an operating system level abstractions and the file system. It also allows the users to use Mapreduce Engine and a Hadoop Distributed File System (HDFS). Hadoop file system gives a name to the rack and its location where the node exists for effective scheduling. This information is then used by the Hadoop application to run the code on the node in which the data exists and then unable to uses the same rack where the data exists and results the reduction of backbone traffic. The same process is also used in Hadoop distributed file system (HDFS) to repeat the data in different racks and also to reduce the impact of rack power outage. Multiple worker nodes and single master nodes are used in the small Hadoop

cluster. Data, Job tracker, Task tracker and the name node are the parts of master node.

HADOOP DISTRIBUTED FILE SYSTEM

Hadoop Distributed File System (HDFS) is the distributed file system that is suitable with very large scale bandwidth. It uses Java application programming interface and also the various shell commands to store the data. Hadoop uses single name node and also the cluster of data nodes that are related with the redundancy options. Every data node is served by the block protocol that is related to HDFS. It uses the TCP/IP sockets for communication purpose and for the communication of clients with each other. Remote procedure calls (RPC) is used. Files are maintained from gigabytes to terabytes.it does not requires the redundant array of independent risks like other file system requires for the data to be stored on hosts because data is replicated on the multiple hosts.

Challenges in Big Data

Big Data is a huge collection of structured and unstructured data therefore it is a challenge to manage the data by using conventional database and software techniques. There are various challenges like capturing of data, analyzing data, sharing of data, searching a particular data, to store data, transfer of data, visualization and privacy of data. According to Akerkar and Zicari, 2014 on the basis of data life cycle the main challenges of Big Data are divided into three main categories:

- **Data Challenge:** This challenge is related to the characteristics of the big data like volume, variety, velocity, veracity, volatility, quality and discovery.
- **Process Challenge:** This challenge is related to series of techniques like how data is to be captured, how data is to be integrated, how data is to be transformed, how the perfect model is to be selected for the analysis and how the results are to be provided.
- **Management Challenge:** This challenge is associated with the data privacy, data security, data governance, garbage mining and various ethical aspects related to data.

Solutions Using Hadoop

The following are the solutions to the above mentioned challenges:

HADOOP Ecosystem

Hadoop Ecosystem is also known as Apache Hadoop Ecosystem that enables the approaches of the MapReduce paradigm at distant conditions and also gives the end-to-end solutions to the various problems. It is vast in nature and grown rapidly due to the contribution of different organizations to this open source

initiative. The Apache Hadoop Ecosystem consists of the following:

- **Hbase**: It is an open source distributed management system that is based on the BigTable by Google. It allows the users to read or write data directly from Hadoop distributed file system. It always runs on the top of Hadoop distributed file system (HDFS). As it does not supports Structured Query Language (SQL) therefore it is also known as NOSQL database. By default it depends on the instance of ZooKeeper.
- **ZooKeeper**: It is the service that contains two nodes, master node and slave node to store the configuration information and also maintains coordination among various nodes.
- **HCatalog:** It is used to store the metadata and then generates the tables for massive amount of data. It is always on the top of Apache Hadoop Ecosystem and helps all the software's to run effectively and store their schemas in Hadoop distributed file system (HDFS).
- **Hive**: It is a platform in the Hadoop Ecosystem that provides data warehouse capabilities and creates its own query language that is known as HiveQL that is compiled by using MapReduce and then implements user-defined functions. It is based on following data structures: Tables that are related to Hadoop distributed file system (HDFS) directories, Partitions in which these directories are distributed are known as Buckets.
- **Pig**: It the framework that develops a high-level scripting language known as Pig Latin. This generated language is used to execute the MapReduce on Hadoop framework. It have better potential data format than Hive. It also have their own data type and map that are used to represents the semi structured data including JavaScript Object Notation (JSON) and extensible Markup Language (XML).
- **Mahout:** It is also an open source library that is used for machine learning and data mining. It is categorized into four main groups they are Clustering, Collective Filtering, Categorization and Mining of the Parallel Patterns. The library of Mahout is belongs to the group that can be optimized to execute the MapReduce framework over Hadoop distributed file system (HDFS).
- **Oozie:** In Hadoop Ecosystem the flow of various jobs are managed, coordinated and then executed at this level. It is also integrated with other Apache Hadoop

frameworks such as Java MapReduce, Streaming MapReduce Pig, Hive and Distcp Sqoop. These actions are combined by Oozie and then arranged the tasks by using Directed Acyclic Graph (DAG).

- **Flume**: It is used to aggregate and transfer huge amount of data inside and outside of Hadoop system. It uses source and sinks as channels. Sources includes system logs, Avro and files whereas Sinks includes HBase and HDFS.
- **Sqoop**: At this level of ecosystem various data sources are imported and exported and then also builds the connection between non Hadoop data sources and HDFS.
- **Ambari:** It gives the set of various jobs to monitor the complications of the Hadoop framework. It also provides the different features like job performance, installation wizard, provisioning and management of Hadoop cluster, system alerts and metrics etc.

HADOOP MapReduce

It is the core of Hadoop and a model which facilitates the mass scalability across various servers in the Hadoop cluster. In this cluster, each server consists of a set of internal disk drives. To increase the performance various assignments are assigned to the servers that processed the data and then stores that data. Processing of data is scheduled according to the cluster of nodes. In Hadoop programs MapReduce performs the works in two parts that is map and reduce. Map is programming task that is divided into various subtasks and that are shared by multiple machines for processing. Here all the stages are stateless and executed independently in distributed environment. Before the reduce task phase starts all the map tasks should be finished. Reduce task is used for multiple pairs. It combines the one or more map tasks to create a result for the reduce task. It also allows to handle the data that is too large to accommodate in the memory. MapReduce also have two entities that are used to finish the process are job tracker and multiple task tracker. Job tracker is used as the model that completes the execution of all scheduled jobs and coordinates with their activity when the multiple tasks are executed on different nodes. It also reschedule the jobs on different task tracker in case of task failure whereas the multiple task tracker acts as a slave and forward the progress to the job tracker regarding its completion.

R Data Language

R is a data language is used for statistical computing and visualization. It was firstly developed at Bell Labs on S project. It is an object

oriented programming language that is used to diverse functions for the analysis of data. This package extends the ability for computing and graphics. It is very useful for the analysis of big data because big data contains text data mostly. R data language is the combination of two modules that is Basic also known as R base which is used for visualization, descriptive statistics, statistical model, data manipulation and data structure and Advanced also known as R package which is used for E1071, clustering, Text Mining (TM), Social Network Analysis (SNA) graph and Sampling. This combination is useful to most of the methods of data science. Data science is known as the study of data that includes all the areas regarding the collection of data, transformation of data, architecture, storage, analysis, visualization, and deployment. For the composition of data firstly data is collected from the various sources that contains statistical sampling for big data and if there is difficulty to analyze that data there should be the legacy of data is to be performed by sampling. Secondly, the data is stored and managed by using database system and cloud computing and then data architecture is to be created like data model and data structure. Based on this data structure analysis of data is to be performed.

Challenges in R Data Language

R Data Language have a few difficulties like Memory Management, Speed and Efficiency. Apart from this many users are using R language instead of other languages but still R is considered as odd language while working with enormous informational collections the structure of the language can prompt a few issues. Information must be stored in the physical memory. Security was not an in-built element to the R language. Additionally, R can't be implanted in a Web browser and it can't be utilized as Web-like or Internet-like applications. It was essentially beside difficult to utilize R as back-end server to perform computations because of absence of security over the Web. For quite a while, there was not a great deal of intuitiveness in the language.

Strategies using R Data Language

Data can be tackled in R Data Language by using these different strategies:

- **Sampling:** If the information is enormous in size and to be examined totally then its size can be diminished by the way toward inspecting that lead to different fulfilling models particularly when the sample is kind of numbers and little in extent to the full informational collection that isn't biased as well.

- **Bigger Hardware:** R data language keeps all articles in memory, yet it is an issue if the information is excessively huge. One of the least complex path is to expand the machine's memory.
- **Store data on hard disk and to analyze it piece wise:** There are different packages accessible to abstain from putting away information into the memory. Rather, objects are put away on hard disk they are examined piece wise. The piecing likewise prompts parallelization normally, if the calculations permit parallel investigation of these pieces.
- **To combine higher performing programming languages like C++ or Java:** There is an option is to combine various programming languages since it the refined method to manage information either with R or with other languages related to their performance.

IV CONCLUSION

Big Data is becoming popular for logical information examine and for business applications. This is also becoming source for the knowledge extraction from the large amount of data that process is known as Big Data Analytics it is a method for breaking down information from these tremendous volumes of data and then resourceful data is driven to maximize the resources. The point of this investigation is to discuss about the different advancements that work together as a Big Data Analytics framework that can help anticipate future volumes, gain experiences, take proactive activities, and offer approach to better vital basic leadership. Further this paper investigates the effect of Big Data Analytics that is utilized to improve its upper hand by utilizing a lot of information calculations for big data sets Hadoop and MapReduce techniques are used.

Hadoop is the most commonly used technique for storing and managing big data and to analyze the tasks involved in large distributed systems that are flexible and scalable. MapReduce is also commonly used for the efficient analysis of data. For the creation of powerful and reliable statistical model R data language is used. Through R data language there is proper analysis of data is to be done.

REFERENCES

[1] Hrishikesh Karambelkar, "Scaling Big Data with Hadoop and Solr", Birmingham B3 2PB, UK, 2013.

[2] Akerkar, R. (2014). "Big data computing". Florida, USA: CRC Press, Taylor & Francis Group.

[3] Zicari, R. V. (2014). "Big Data: Challenges and Opportunities" (2014) In R. (Ed.), *Big data computing* (pp. 103–128). Florida, USA: CRC Press, Taylor & Francis Group.

[4] J. Stanton, "Introduction to Data Science", Syracuse University, (2013).

[4] Jasleen Kaur Bains. "Big Data Analytics in Healthcare- Its Benefits, Phases and Challenges". *International Journal of Advanced Research in Computer Science and Software Engineering*. 2016; 6(4): 430–435.

[5] Nada Elgendy, Ahmed Elragal. "Big data analytics: A literature review paper." *Springer International Publishing*. 2014; 8:214–227.

[6] P Harshawardhan S. Bhosale, Prof. Devendra. Gadekar, "A Review paper on Bigdata and Hadoop" *International Journal of Scientific and Research Publications*, Volume 4, Issue 10, October 2014.

[7] Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST),* May 2010, pp. 1–10.

[8] J. Dean and S. Ghemawat,"Mapreduce: simplified data processing on large clusters,"Communications of the ACM, vol. 51, no. 1, pp.107–113, 2008.

[9] Song, Z., Kusiak, "A.: Optimizing Product Configurations with a Data Mining Approach". *International Journal of Production Research* 47(7), 1733–1751 (2009)

[10] Adams, M.N.: "Perspectives on Data Mining" *International Journal of Market Research* 52(1), 11–19 (2010)

[11] "Big Data Analytics What it is and why it matters", SAS, accessed 14 May 2018, https://www.sas.com/en_us/insights/analytics/big-data-analytics.html

[12] Vangie Beal. (2018), "Big Data Analytics", accessed 10 May 2018, <https://www.webopedia.com/TERM/B/big_data_analytics.html>

[13] Irwin King, Michael R. Lyu and Haiqin Yang, "online learning for big data analytics". 2013. p. 1–116.

[14] Waller, M. A., & Fawcett, S. E. (2013). "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management". Journal of Business Logistics, 34(2), 77–84.

[15] Fan, J., Han, F., & Liu, H. (2014). "Challenges of big data analysis". *National Science Review*, 1(2), 293–314.

[16] "R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing", Vienna, Austria, http://www.R-project.org, (2015).

[17] J. Stanton, "Introduction to Data Science", Syracuse University, (2013).

[18] S. Jun, S. Leeand J. Ryu, "A Divided Regression Analysis for Big Data", *International Journal of Software Engineering and Its Applications*, vol. 9, no. 5, (2015), pp. 21-32.

[19] Wikipedia, the free encyclopedia, http://en.wikipedia.org, (2014).

[20] Naresh Babau, Suneetha Mane. A comprehensive survey of big data analytics and techniques. *IJCST*. 2016; 14(10):177–184.